

EVALUATING ``HUMAN + ADVISORY COMPUTER'' SYSTEMS: A CASE STUDY

Andrey A. Povyakalo
Centre for Software
Reliability (CSR)
City University, London
andrey@csr.city.ac.uk

Eugenio Alberdi
Centre for Software
Reliability (CSR)
City University, London
e.alberdi@csr.city.ac.uk

Lorenzo Strigini
Centre for Software
Reliability (CSR)
City University, London
l.strigini@csr.city.ac.uk

Peter Ayton
Psychology Dept
City University, London
p.ayton@city.ac.uk

ABSTRACT

We studied the impact of a Computer Aided Detection (CAD) tool on human's decisions in mammography. We used data from an independent clinical trial, which compared the average performance of breast screening professionals with and without CAD. Standard analyses of these data showed no statistically significant effect of CAD's output on humans' cancer detection rate. We conducted statistical modelling and supplementary analyses of these data focusing on the role of the correctness of the computer output and the variation of the difficulty of decisions. These analyses provided counter-intuitive results: e.g., incorrect output from CAD was likely to increase the difficulty of difficult-to-detect cancers; at the same time, correct computer output systematically improved human decisions for easy-to-detect cancers. Our findings indicate that, even when its average impact in a trial is zero, computer support can have non-obvious systematic effects on human performance which are worth considering during system design and evaluation.

Prior permission is required to copy or otherwise to republish all or part of this work in any form or medium, unless for personal or classroom use and without commercial advantage. Any copies that are made must bear this notice and the full citation on the first page

Proceedings Volume 2 of the Conference
HCI 2004: Design for Life
Leeds, UK, September 2004.
© 2004 British HCI Group

Keywords

Human-computer interaction, dependability, decision making, mammography, clinical trial, statistical analysis

1. INTRODUCTION

Computer support of human decision making is increasingly common in many areas, including critical areas like health care.

The intended role of automation is often purely auxiliary: to compensate for the risk that humans miss some information (e.g. when fatigued or overloaded), or to make the process of human apprehension of all important information more efficient. The human user retains the authority and responsibility for decisions.

The assumption that "the computer support may improve things but not make them worse" would be very helpful in decisions about deploying a system. So, it is important to ask whether the assumption holds in a specific case; in which cases it is likely not to hold; and how to judge e.g. how unreliable a support must be for it to stop helping and start damaging human decision making.

The work described in this paper was conducted as part of a UK-funded interdisciplinary, collaborative project, DIRC, which deals with the dependability (e.g., reliability, security, availability, etc.) of computer based systems: systems encompassing not only computer software and hardware but also their users and the social and physical context of computer use.

2. BACKGROUND

We consider the use of a particular Computer Aided Detection (CAD) tool to assist the interpretation of mammograms (breast X-rays) in screening for breast cancer.

In CAD, a computer aid is to recognise and mark (“prompt”) regions of interest (ROI) on a digitized mammogram, to prevent clinicians from overlooking them. CAD is not meant to be a diagnostic tool, in the sense that it only marks ROIs, which should be subsequently classified by a human expert (“reader”).

In accordance with the prescribed procedure for the CAD tool under consideration, a human reader looks at a mammogram and interprets it as usual, then activates the CAD tool and looks at a digitized image of the mammogram with marked ROIs, checks whether he/she has overlooked any features with diagnostic value and, if appropriate, revises his/her original assessment.

The manufacturers of the tool claim that, when the CAD tool is used as recommended, the potential for missing lesions is not increased over routine screening.

Our analysis uses data from an independent clinical trial, which evaluated the CAD tool under consideration[5]. In this trial 50 human readers looked at a mixture of mammograms, representing 120 normal cases and 60 cancers, in two conditions: with and without CAD support (“prompted” and “unprompted” conditions). Each reader went through all the cases in both conditions to decide whether or not each case should be recalled for further investigation.

This study found no statistically significant effect of CAD prompts on the performance of these readers in detecting cancers.

3. METHOD

The analyses we describe in this paper were inspired by studies of the dependability of software systems with diverse redundancy [2]. These studies have shown that the variation of the difficulty of input cases for system components (its variation across the possible cases, and the co-variation between the difficulty for different system components) affects the dependability of the overall system substantially. These studies also indicate that focusing on the average probabilities of the failures of the components, as well as assuming independence between their failures, can be misleading. We explored the possibility that some of the conclusions from those studies could also apply to this case, in which the redundant system is made up of the reader and the CAD tool.

Therefore we focused our analyses on the role of correctness of computer output and the variation of difficulty of cases in determining the probability of a wrong decision by a reader.

The scatter plot in figure 1 represents part of the data from the clinical trial. It shows all cancer cases that we categorize as “non-obvious”, that is, cancers, which were missed by at least one reader in either the prompted or the unprompted condition. Each cancer is represented by a symbol: symbols “c” represent correctly prompted cancers; symbols “w” represent incorrectly prompted cancers. Some symbols represent several cancers.

The horizontal axis represents the fraction of readers who - in the unprompted condition - missed the cancer (we call this number the “difficulty”, d , of that cancer).

The vertical axis presents the same fraction but measured in the “prompted” condition (i.e. with support by the CAD tool).

If there is no impact of prompts on human detection rate, then any deviation of the triangles from the unit slope line can only be interpreted as being caused by statistical noise.

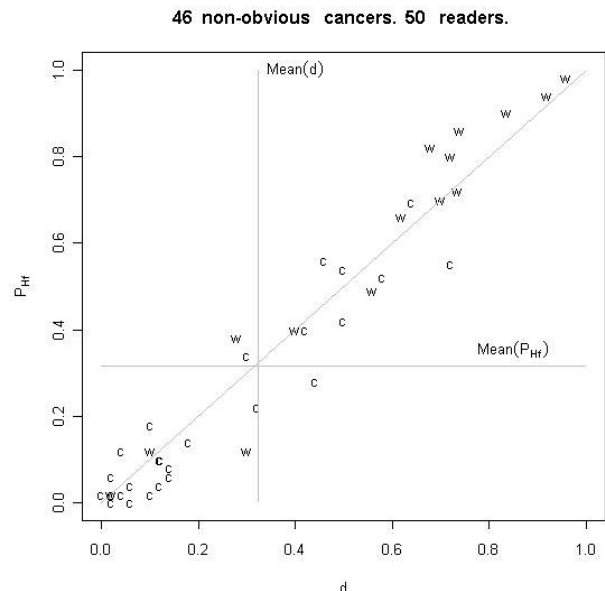


Figure 1: Trial data for non-obvious cancers

Our analyses aimed at eliminating random statistical noise, to estimate properly any systematic effect from CAD output on a decision of any given human reader about any given case. We found that for the given data the best way to do this is the logistic regression (a. k. a. generalized linear model with binary response and logit link, e.g. see [1]).

We must state explicitly that at this stage we do not consider our statistical models as explanatory or predictive. They are just tools for eliminating statistical noise when analysing this particular data set.

Logistic regression allows us to estimate the probability $p(i,j)$ of a reader j missing a cancer i as a function:

$$p(i,j) = F(d_i, np_j, pp_j, wr_i). \quad (1)$$

where the variables are: case difficulty d_i ; rate of missed cancers for a reader np_j ; rate of false alarms for a reader

pp_j ; and wr_i , an indicator of whether CAD marks the case i wrongly.

Logistic regression assumes independent observations of a binary response (failure/success); it does not assume normal error with the same variance for all observations and it always assures

$$0 \leq p_{ij} \leq 1.$$

In our exploratory analyses, instead of fitting a pre-selected statistical model to the data, we extracted a model from the data by stepwise regression using the Akaike Information Criterion (AIC, e.g. see [3]).

The procedure starts with the 'empty' model, in which the right-hand side of equation (1) contains a single, constant term. Then, given a certain collection of possible model terms, at each step it adds the most informative term to the model and removes a term when it becomes noninformative, until no term can be added or removed without increasing the Akaike information criterion (AIC):

$$AIC = Deviance +$$

$$2 \cdot \text{number of parameters} + \text{const}$$

(Deviance is a measure of discrepancy between the observed responses and those estimated by the model [1])

AIC resolves the trade-off between model accuracy and model complexity (which may lead to over-fitting). It penalizes the addition of every new term to the model. A new term may be added to the model only when the resulting reduction of deviance is greater than the penalty for adding the term.

At the end the procedure yielded a model, where all covariates (predictors): d_i , np_j , pp_j , wr_i are significant. It also contains significant nonlinear terms.

The analysis of deviance indicated the adequacy of the whole model, i.e. usual random deviance caused by the 'normative' statistical noise exceeds the observed value with the substantial probability.

4. RESULTS

The model obtained from the above procedure implies an estimated probability of wrong decision by the readers, for any given cancer. The estimated probabilities $p_{cor}(d)$, $p_{wr}(d)$ of a randomly selected reader missing a cancer with difficulty d , given a correct and respectively given a wrong CAD output, are given by the averages (over the m readers):

$$p_{cor}(d) = \frac{1}{m} \sum_{j=1}^m F(d, np_j, pp_j, 0)$$

$$p_{wr}(d) = \frac{1}{m} \sum_{j=1}^m F(d, np_j, pp_j, 1)$$

With the same method one can obtain the function

$$p_0(i, j) = F_0(d_i, np_j, pp_j),$$

to get the estimate for all correctly and wrongly prompted cases together:

$$p(d) = \frac{1}{m} \sum_{j=1}^m F_0(d, np_j, pp_j)$$

In the plot in Figure 2, the horizontal axis represents the difficulty d . The vertical axis shows the differences $P-d$ between the estimated probability P of a given cancer being missed by a randomly selected reader supported by CAD and the cancer's difficulty d . So, a point below the horizontal line $P-d = 0$ indicates a cancer for which CAD appears to reduce the rate of reader errors.

The symbols "w" and "c" show the observed values of d and $P-d$ for non-obvious cancers with correct CAD output (symbols "c") and wrong CAD output (symbols "w").

The curves represent the functions: $p_{wr}(d) - d$ (dashed curve), $p_{cor}(d) - d$ (dotted-dashed curve) and $p(d) - d$ (solid curve).

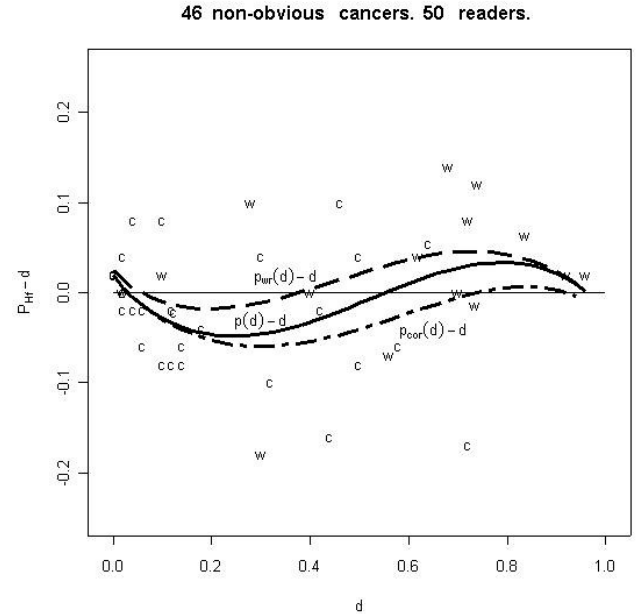


Figure 2: Effects of case difficulty and prompt correctness

5. CONCLUSIONS

These exploratory analyses suggest that in the context under consideration:

- correct computer output is likely to improve readers' decisions for cases of any difficulty d : see figure 2;
- readers are likely to tolerate wrong computer output for easy cases (with $d < 0.4$); however, for the difficult cases ($d > 0.4$) wrong computer output increases the rate of missed cancers: see figure 2;
- as a result of the combination of these two effects, CAD systematically decreases the rate of cancers missed by a randomly selected reader for easy

cancers (with $d < 0.6$) and increases this rate for the difficult cancers (with $d > 0.6$): see figure 2;

- the total average effect on the rate of missed cancers (in the set used in this trial) is close to 0, as reported by previous analyses [2];

The indication that CAD, at least in this experimental situation, systematically improved or damaged the performance of readers in a way that depended on the difficulty of the individual case, seems important. It implies that even if the average effect of CAD detected in a clinical trial was zero, if CAD were used in practice with a different mix of patients from that used in the trial (which in any case had to be unrealistic, from certain viewpoints, due to needs of experiment design) there may be non-zero (either useful or harmful) average effects. Secondly, it implies that a way of improving the effect of CAD on the effectiveness of cancer screening may be in seeking the causes of these systematic effects in the design of a CAD tool and the procedures for its use. More in general, it seems to confirm the usefulness of an analytical approach, as we have proposed [3] which focuses on *variations* in effectiveness across different cases to estimate the effects of possible design improvements. Last, they support a viewpoint that the visible details of the *interface* of a computer system may be a comparatively unimportant part of the *interaction* that determines its effects in supporting human work: in this case, we would conjecture that design choices like the false positive-false negative trade-offs selected by designers for various categories of cases, and the way readers learn about the characteristics of the CAD tool, may be the determinant factors in the effects observed here.

Our bullet list of suggested conclusions is a set of tentative hypotheses derived from the exploratory analyses of a particular data set. They are subject to confirmation by a confirmatory analysis of data from other clinical studies, and to the test of whether the effects observed can be explained by convincing theories of their causes in human behaviour. This is work in progress.

One can apply the same method we used in analyzing false negative rates to analyse the effect of CAD on the rate of false positives (recalled normal cases), and thus reason about the design trade-offs in the design of the whole human-computer system.

ACKNOWLEDGMENTS

This work was supported in part by the U.K. Engineering and Physical Sciences Research Council via the Interdisciplinary Collaboration on the Dependability of Computer Based Systems, "DIRC".

Special thanks go to Paul Taylor, from University College London (UCL) for granting access to the data from the clinical trial (funded by the UK Health Technology Assessment programme) and for providing valuable feedback. We would also like to acknowledge the collaboration of Jo Champness (from UCL) and the readers who participated in the trial.

REFERENCES

- [1] Dobson, A., J., An introduction to generalized linear models. London: Chapman and Hall 1990.
- [2] Littlewood, B., Popov, P. and Strigini, L., Modelling software design diversity - a review, *ACM Computing Surveys*, 33, no. 2, 177-208, (2002).
- [3] Strigini, L., Povyakalo, A. A., Alberdi, E. Human-machine diversity in the use of computerised advisory systems: a case study. In: *Proceedings of DSN 2003, International Conference on Dependable Systems and Networks, San Francisco*, 249-258 (2003).
- [4] Sakamoto, Y., Ishiguro, M., and Kitagawa G. Akaike Information Criterion Statistics. D. Reidel Publishing Company. 1986
- [5] Taylor, P.M., Champness, J., Given-Wilson, R. M., Potts H.W.E. and Johnston K. An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms *British Journal of Radiology*, 77(913): 21-27(2004)